

UNITED STATES PATENT APPLICATION

of

Sanjay Anand

Kyle Brandon

Nk Srinivas

and

Jameel Hyder

for

**METHOD AND COMPUTER PROGRAM PRODUCT FOR OFFLOADING
PROCESSING TASKS FROM SOFTWARE TO HARDWARE**

WORKMAN, NYDEGGER & SEELEY
A PROFESSIONAL CORPORATION
ATTORNEYS AT LAW
1000 EAGLE GATE TOWER
60 EAST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84111

BACKGROUND OF THE INVENTION

1. Cross-Reference to Related Applications

The present application is a continuation-in-part application of United States patent application serial number 09/657,510 filed 7 September 2000 and entitled "METHOD AND COMPUTER PROGRAM PRODUCT FOR OFFLOADING PROCESSING TASKS FROM SOFTWARE TO HARDWARE" which application is incorporated herein by reference in its entirety.

2. The Field of the Invention

The present invention relates generally to methods for increasing the efficiency, speed and/or throughput of a computer system. More specifically, the invention relates to methods for offloading computing tasks that are typically performed by a host processor in software, to a specific hardware component, thereby freeing up host processor resources and increasing the overall efficiency of the computer system.

3. The State of the Art

A functional computer system generally consists of three fundamental components. The first component is the host computer and its associated peripheral hardware components. The host computer typically includes a central processing unit (CPU), which is interconnected via a bus with, for instance, system memory such as RAM or ROM. A system will also include a number of peripheral hardware devices, depending on the functionality needed, such as magnetic or optical disk storage devices, a keyboard or other input device, a display or other output device and communication equipment, such as a modem and/or a network interface card (NIC). Another fundamental computer component

1 is the application software. Such software includes the familiar word processing
2 applications, spread sheet applications, database applications, communications and
3 network applications and so forth.

4 The final component of a modern, functional computer system is an operating
5 system. The computer operating system performs many functions such as allowing a user
6 to initiate execution of an application program. In addition, modern operating systems also
7 provide an interface between application software and the host computer and its peripheral
8 hardware. Thus, while it was once commonplace for an application program to directly
9 access computer system hardware, modern operating systems provide standardized,
10 consistent interfaces that allow user applications to interface with or access computer
11 hardware peripherals in a standardized manner. To provide a consistent interface,
12 operating system architectures are increasingly designed so that there may be several
13 software layers between the actual hardware peripheral and the application program. For
14 example, an application may make a call into the operating system. The operating system,
15 in turn, may utilize the services provided by a hardware device driver layer. The device
16 driver layer would then interface directly with the specific hardware peripheral. A primary
17 advantage of such a layered approach is that layers may be added or replaced without
18 impacting the other layers.

19 As will be appreciated, the complexity and sophistication of such operating
20 systems, application software, and networking and communications continues to increase.
21 This of course results in more functional and useful computer systems. However, this
22 increased functionality is not without a cost. More feature rich operating systems and
23 software applications often result in an increase in the processor overhead as a result of the
24 additional duties that must be performed by a processor/CPU when executing such system

1 functions and/or applications. This phenomenon is especially apparent in connection with
2 particular types of applications, such as network communication-type software
3 applications. With the high bandwidth media that is increasingly prevalent, network
4 speeds often match or exceed the CPU processor speed and memory bandwidth of the host
5 computer. As such, to efficiently communicate over such networks, the CPU utilization
6 and memory bandwidth used of the network-connected host computer must be minimized.

7 In addition, network applications further burden the host processor due to the
8 layered architecture used by most, such as the seven-layer ISO model, or the layered model
9 used by the Windows NT operating system. As is well known, such a model is used to
10 describe the flow of data between the physical connection to the network and the end-user
11 application. The most basic functions, such as putting data bits onto the network cable, are
12 performed at the bottom layers, while functions attending to the details of applications are
13 at the top layers. Essentially, the purpose of each layer is to provide services to the next
14 higher layer, shielding the higher layer from the details of how services are actually
15 implemented. The layers are abstracted in such a way that each layer believes it is
16 communicating with the same layer on the other computer that is being communicated
17 with via the network.

18 As will be appreciated, the various functions that are performed on a data packet as
19 it proceeds between layers can be software intensive, and thus can demand a substantial
20 amount of CPU processor and memory resources. For instance, in the Windows NT
21 networking model, certain functions that are performed on the packet at various layers are
22 extremely CPU intensive, such as packet checksum calculation and verification; encryption
23 and decryption of data (e.g., SSL encryption and IP Security encryption); message digest
24 calculation, TCP segmentation, receive side packet classification, packet filtering to guard

1 against denial of service attacks, and User Datagram Protocol (UDP) send side packet
2 fragmentation. As each of these functions are performed, the resulting demands on the
3 CPU/memory can greatly effect the throughput and performance of the overall computer
4 system.

5 Although software applications and operating system functions are placing greater
6 demands on computer system resources, at the same time the capability, efficiency, and
7 throughput of many computer hardware peripherals -- such as network interface cards
8 (NICs) -- is also increasing. These computer system peripherals are often equipped with a
9 dedicated processor and memory, and typically are capable of performing very
10 sophisticated and complex computing tasks -- tasks that are otherwise performed by the
11 computer system processor in software. For instance, many NICs are capable of
12 independently performing tasks otherwise performed by the CPU in software at an
13 appropriate network layer, such as checksum calculation/verification; data
14 encryption/decryption; message digest calculation; TCP or UDP segmentation; receive side
15 packet classification; packet filtering to guard against denial of service attacks; and others.
16 As such, there is an advantage in offloading such CPU intensive task to a peripheral
17 hardware device. This would reduce processor utilization and memory bandwidth usage in
18 the host computer, and thereby increase the efficiency, speed and throughput of the overall
19 system.

20 However, the processing capabilities of different peripheral devices vary widely.
21 Thus, there needs to be an efficient method by which a computer system/operating system
22 can identify the processing capabilities of such peripheral devices, and then assign and
23 offload specific processing tasks to the device when needed. Also, it would be desirable if
24 the tasks could be identified and assigned dynamically, depending on the then current

1 needs of the processor. This would allow the computer system processor to take advantage
2 of the capabilities of a hardware peripheral on an as-needed basis.
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

SUMMARY AND OBJECTS OF THE INVENTION

The foregoing problems in the prior state of the art have been successfully overcome by the present invention, which is directed to a system and method for offloading functions and tasks that were previously performed at a processor-software level, to an appropriate hardware peripheral connected to the computer system. The invention is particularly useful in connection with the offloading of tasks to network interface card (NIC) peripheral devices, which can often perform many of the tasks otherwise performed by the computer CPU in software.

In a preferred embodiment of the invention, a software implemented method and protocol is provided that allows, for instance, the operating system (OS) to "query" the device drivers (often referred to as "MAC" drivers) of any hardware peripherals (such as a NIC) that are connected to the computer system. The various device drivers each respond by identifying their respective hardware peripheral's processing capabilities, referred to herein as "task offload capabilities." In the preferred embodiment, once the task offload capabilities of each particular peripheral have been identified, the OS can then enable selected peripherals to perform certain tasks that could potentially be used by the OS. The OS can thereafter request that a peripheral perform the previously enabled task, or tasks, in a dynamic, as-needed basis, depending on the then current processing needs of the computer system.

While this general inventive concept would be applicable to other application or operating system environments, embodiments of the current invention are described herein as being implemented and utilized in connection with the layered networking model of Windows NT. Of course, the invention could be implemented in connection with essentially any similar type of architecture for managing and controlling network

1 communications. Specifically, the invention provides the ability to offload tasks or
2 functions that are typically performed on a network packet at, for instance, the various
3 network layers, and which typically require dedicated CPU and memory resources. These
4 offloaded tasks can instead be optionally performed by the hardware peripheral that
5 provides the actual physical communications channel to the network -- the NIC. For
6 instance, rather than perform certain of the CPU intensive operations on the data packet as
7 it passes through the respective network layers -- e.g. checksum calculation/verification,
8 encryption/decryption, message digest calculation; TCP or UDP segmentation, receive side
9 packet classification; packet filtering to guard against denial of service attacks; -- those
10 tasks can instead be offloaded and performed at the NIC hardware.

11 In a preferred embodiment of the present invention, in the Windows NT layered
12 networking architecture, a transport protocol driver, or transport, is implemented with an
13 appropriate program method so as to be capable of querying each of the device driver(s)
14 associated with the corresponding NIC(s) connected to the computer. Each queried device
15 driver is similarly implemented so as to be capable of responding by identifying its specific
16 processing, or "task offload" capabilities. In a preferred embodiment, once the task offload
17 capabilities of each individual peripheral device have been identified, the transport sets
18 which of those specific capabilities are to be enabled. This essentially informs the
19 peripheral device what type of tasks it should expect to perform during subsequent
20 transmissions and/or receptions of data packets. Thereafter, the transport is able to take
21 advantage of the enabled capabilities of a peripheral device on an as-needed basis.
22 Preferably, the enabled functions are invoked via appropriate data that is appended to the
23 actual data packet destined for the network channel. In this way, tasks can be offloaded
24 dynamically, and more than one task can be offloaded at a time.

1 Thus, before a network packet is to be sent to a particular lower level device driver
2 (e.g., residing at the MAC sublayer in a Windows NT environment), the transport will first
3 determine what the capabilities of the corresponding NIC are. If capable of a particular
4 function or functions, the transport enables the desired functions. If during subsequent
5 packet transmissions the transport desires that a particular task be offloaded to hardware, it
6 can dynamically append information to the packet that signifies that the desired function(s)
7 should be performed on that packet at the NIC hardware. For instance, the transport will
8 set a data flag in the data packet, thereby notifying the corresponding device driver that the
9 NIC should calculate and append a checksum to that outgoing packet. The
10 hardware/software on the corresponding NIC will then handle this particular packet
11 processing on its own, without any intervention or assistance from the system CPU. The
12 system processor is thus freed up to perform other processing tasks, and the overall
13 efficiency and throughput of the system is improved.

14 As noted, in a preferred embodiment, tasks are downloaded dynamically. That is,
15 the capability of the NIC can be selectively used on a per-packet basis, depending on the
16 then current needs of the computer system. Moreover, since tasks that were previously
17 performed at various levels of the network stack are now performed at a single point -- the
18 NIC itself -- the approach is more tightly integrated and efficient, further improving the
19 throughput of the entire system. Preferably, embodiments of the current invention provide
20 the transport with the ability to "batch" operations, *i.e.*, offload multiple tasks to a single
21 NIC. For instance, a single NIC can perform both checksumming, and encryption on a
22 packet, thereby eliminating multiple CPU cycles that would have otherwise been needed if
23 the same functions were implemented at the respective network layers in software.
24

1 Accordingly, it is a primary object of this invention to provide a system and method
2 for offloading computing tasks from a computer system processor to a hardware peripheral
3 connected to the computer system. It is another object of this invention to provide a
4 system and method for identifying the processing capabilities of individual peripherals. It
5 is still another object of the present invention to provide a system and method for
6 offloading tasks that can be effectively used in connection with a layered network
7 architecture, whereby tasks that are typically performed at various layers in the network are
8 instead offloaded to the appropriate network interface card (NIC). A related object of the
9 invention is to provide a system and method in which computing tasks can be offloaded in
10 a dynamic, as-needed basis, depending on the then current processing state of the computer
11 processor. Yet another object of the present invention is to provide a system and method
12 in which multiple tasks can be batched together, and then offloaded to a single peripheral
13 device, such as a NIC.

14 Additional objects and advantages of the invention will be set forth in the
15 description which follows, and in part will be obvious from the description, or may be
16 learned by the practice of the invention. The objects and advantages of the invention may
17 be realized and obtained by means of the instruments and combinations particularly
18 pointed out in the appended claims. These and other objects and features of the present
19 invention will become more fully apparent from the following description and appended
20 claims, or may be learned by the practice of the invention as set forth hereinafter.
21
22
23
24

BRIEF DESCRIPTION OF THE DRAWINGS

In order that the manner in which the above-recited and other advantages and objects of the invention are obtained, a more particular description of the invention briefly described above will be rendered by reference to specific embodiments thereof which are illustrated in the appended drawings. Understanding that these drawings depict only typical embodiments of the invention and are not, therefore, to be considered to be limiting of its scope, the invention will be described and explained with additional specificity and detail through the use of the accompanying drawings in which:

Figure 1 is a diagram representing an example of a typical computer system and attached peripheral devices that can be used in connection with the present invention;

Figure 2 is a diagram illustrating some of the functional components present in a layered network architecture;

Figure 3 is a functional block diagram illustrating the flow of a data packet through program components in accordance with one presently preferred embodiment of the invention;

Figure 4 is a diagram illustrating one presently preferred embodiment of the data packet and the packet extension;

Figure 5 is flow chart illustrating one presently preferred embodiment of the program steps used to offload tasks on per-packet basis; and

Figure 6 is a flow chart illustrating one presently preferred embodiment of the program steps used to query and set the task offload capabilities of a peripheral device.

DETAILED DESCRIPTION OF THE INVENTION

The invention is described below by using diagrams to illustrate either the structure or processing of embodiments used to implement the system and method of the present invention. Using the diagrams in this manner to present the invention should not be construed as limiting of its scope.

The present invention contemplates both methods and systems for offloading processing tasks from a host computer, such as a personal computer, to computer connected hardware peripherals, such as a network interface card (NIC). Figure 1 and the following discussion are intended to provide a brief, general description of a suitable computing environment in which the invention may be implemented. Although not required, embodiments of the invention will be described in the general context of computer-executable instructions, such as program modules, being executed by a personal computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Moreover, those skilled in the art will appreciate that the invention may be practiced with other computer system configurations, including hand-held devices, multiprocessor systems, microprocessor-based or programmable consumer electronics, network PCs, minicomputers, mainframe computers, and the like. Embodiments of the invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

With reference to Figure 1, an exemplary system for implementing the invention includes a general purpose computing device in the form of a conventional personal

1 computer 20, including a processing unit 21 (sometimes referred to as the CPU), a system
2 memory 22, and a system bus 23 that couples various system components including the
3 system memory to the processing unit 21. The system bus 23 may be any of several types
4 of bus structures including a memory bus or memory controller, a peripheral bus, and a
5 local bus using any of a variety of bus architectures. The system memory includes read
6 only memory (ROM) 24 and random access memory (RAM) 25. A basic input/output
7 system 26 (BIOS), containing the basic routines that help to transfer information between
8 elements within the personal computer 20, such as during start-up, is stored in ROM 24.
9 The personal computer 20 further includes a variety of peripheral hardware devices, such
10 as hard disk drive 27 for reading from and writing to a hard disk, not shown, a magnetic
11 disk drive 28 for reading from or writing to a removable magnetic disk 29, and an optical
12 disk drive 30 for reading from or writing to removable optical disk 31 such as a CD ROM
13 or other optical media. The hard disk drive 27, magnetic disk drive 28, and optical disk
14 drive 30 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic
15 disk drive-interface 33, and an optical drive interface 34, respectively. The drives and their
16 associated computer-readable media provide nonvolatile storage of computer readable
17 instructions, data structures, program modules and other data for the personal computer 20.
18 Although the exemplary environment described herein employs a hard disk, a removable
19 magnetic disk 29 and a removable optical disk 31, it should be appreciated by those skilled
20 in the art that other types of computer readable media which can store data that is
21 accessible by a computer, such as magnetic cassettes, flash memory cards, digital video
22 disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROM),
23 and the like, may also be used in the exemplary operating environment.

1 A number of program modules may be stored on the hard disk, magnetic disk 29,
2 optical disk 31, ROM 24 or RAM 25, including an operating system 35, one or more
3 application programs 36, other program modules 37, and program data 38. A user may
4 enter commands and information into the personal computer 20 through input devices such
5 as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a
6 microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input
7 devices are often connected to the processing unit 21 through a peripheral hardware device
8 such as a serial port interface 46 that is coupled to the system bus, but may be connected
9 by other interfaces, such as a parallel port, game port or a universal serial bus (USB). A
10 monitor 47 or other type of display device is also connected to the system bus 23 via a
11 peripheral hardware interface device, such as a video adapter 48. In addition to the
12 monitor, personal computers typically include other peripheral output devices (not shown),
13 such as speakers and printers.

14 The personal computer 20 may operate in a networked environment using logical
15 connections to one or more remote computers, such as a remote computer 49. The remote
16 computer 49 may be another personal computer, a server, a router, a network PC, a peer
17 device or other common network node, and typically includes many or all of the elements
18 described above relative to the personal computer 20, although only a memory storage
19 device 50 has been illustrated in Figure 1. The logical connections depicted in Figure 1
20 include a local area network (LAN) 51 and a wide area network (WAN) 52. Such
21 networking environments are commonplace in offices enterprise-wide computer networks,
22 intranets and the Internet.

23 When used in a LAN networking environment, the personal computer 20 is
24 connected to the local network 51 through a peripheral hardware device often referred to as

1 a network interface card (NIC) or adapter 53. When used in a WAN networking
2 environment, the personal computer 20 typically includes a modem 54 or other means for
3 establishing communications over the wide area network 52, such as the Internet. The
4 modem 54, which may be internal or external, and typically is connected to the system bus
5 23 via the serial port interface 46. In a networked environment, program modules depicted
6 relative to the personal computer 20, or portions thereof, may be stored in the remote
7 memory storage device 50. It will be appreciated that the network connections shown are
8 exemplary and other means of establishing a communications link between the computers
9 may be used.

10 Embodiments within the scope of the present invention also include computer
11 readable media having executable instructions. Such computer readable media can be any
12 available media which can be accessed by a general purpose or special purpose computer.
13 By way of example, and not limitation, such computer readable media can comprise
14 physical storage media such as RAM, ROM, EEPROM, CD-ROM or other optical disk
15 storage, magnetic disk storage or other magnetic storage devices, or any other medium
16 which can be used to store the desired executable instructions and which can be accessed
17 by a general purpose or special purpose computer.

18 When information is transferred or provided over a network or another
19 communications connection (either hardwired, wireless, or a combination of hardwired or
20 wireless) to a computer, the computer properly views the connection as a computer-
21 readable medium. Thus, any such a connection is properly termed a computer-readable
22 medium. Combinations of the above should also be included within the scope of computer
23 readable media. Executable instructions comprise, for example, instructions and data
24 which cause a general purpose computer, special purpose computer, or special purpose

1 processing device to perform a certain function or group of functions. Finally,
2 embodiments within the scope of the present invention comprise a computer readable
3 medium having a plurality of data fields stored thereon that represent a data structure.

4 Embodiments of the present invention are directed to providing the ability to
5 reducing the processing overhead and memory usage of a processing unit 21. This is
6 accomplished by offloading particular computing tasks, which are accomplished for
7 instance by way of an operating system, application programs and/or other program
8 modules that are executing on the processing unit/CPU 21, to an appropriate peripheral
9 hardware device connected to the computer system 20. Many such peripheral devices are
10 increasingly equipped with dedicated processors and memory, and are fully capable of
11 performing many of the same tasks that are typically accomplished solely by the CPU 21.
12 Examples of such devices could include, for instance, network interface cards (53 in Figure
13 1); disk drive interface cards (*e.g.*, 32, 33, 34 in Figure 1); small computer system interface
14 (SCSI) devices; intelligent serial interface cards; or application specific peripherals, such
15 as devices for the encryption/decryption of data.

16 While the general inventive concepts discussed herein could be used to offload
17 computing tasks in connection with any of the above peripheral hardware devices, the
18 invention will be described with respect to an example of one presently preferred
19 embodiment, wherein computing tasks are offloaded to a network communications device,
20 such as NIC 53 illustrated in Figure 1. More particularly, illustrative embodiments are
21 discussed as being implemented in the networking environment and architecture of the
22 Windows NT operating system available from Microsoft Corporation. However, it will be
23 appreciated that while specific reference is made to Windows NT concepts and
24 terminology, those skilled in the art will recognize that many, if not most, operating

1 systems and networking architectures share similarities relevant to the environment of the
2 present invention.

3 In order to more fully understand the context of the present invention, reference is
4 next made to Figure 2, which illustrates a simplified diagram of certain of the components
5 that make up the Windows NT networking model. For purposes of illustration, the OSI
6 layers that correspond to the various Windows NT components are also shown. At the
7 bottom layer, corresponding to the physical layer in the OSI model, resides the actual NICs
8 (sometimes referred to as network cards, or network adapters) 100-104. NICs are the
9 hardware devices that provide the physical interconnection with the physical medium (the
10 network cable), and the transmission of the signals that carry the data generated by all the
11 higher level layers, in accordance with the particular network topology. As noted above,
12 many NICs are equipped with a dedicated processor and memory, and are capable of
13 performing additional sophisticated computing tasks -- including tasks that may otherwise
14 be handled by the host processor CPU. NICs can be physically implemented as a printed
15 circuit board card that is positioned within a slot in the computer, as a PCMCIA type card
16 that is placed within a PCMCIA-compliant slot, as a dedicated chip positioned within the
17 computer chassis on the mother board, or in any other suitable matter.

18 Each NIC is logically interconnected with the Windows NT networking model, as
19 is schematically represented by bidirectional lines 108-112, via a corresponding network
20 driver 116-120. Network drivers reside in the MAC sublayer of the network model, and
21 link Windows NT to the physical network channel via the corresponding NICs. Each
22 driver, typically implemented as a software component provided by the vendor of the
23 corresponding NIC, is responsible for sending and receiving packets over its corresponding
24 network connection and for managing the NIC on behalf of the operating system. Each

1 driver also starts I/O on the corresponding NIC and receives interrupts from them, and
2 calls upward to protocol drivers to notify them of its completion of an outbound data
3 transfer. Also, the device driver will be responsible for invoking, controlling and/or
4 monitoring any of the additional processing capabilities of the corresponding NIC.

5 In some environments the driver component is written so as to implement a single
6 specific network protocol, such as TCP/IP or XNS. The basic invention of the present
7 invention described and claimed herein would be applicable to such an environment. For
8 purposes of illustration however, the present invention is described in connection with the
9 Windows NT network architecture, in which an interface and environment called the
10 network driver interface specification (NDIS) is provided. The NDIS interface is
11 functionally illustrated in Figure 2 at 126. NDIS shields each of the network drivers 116-
12 120 from the details of various transport protocols (examples of which are shown at 128-
13 134), and vice versa. More particularly, NDIS describes the interface by which one or
14 multiple NIC drivers (116-120) communicate with one or multiple underlying NICs (100-
15 104), one or multiple overlying transport protocol drivers, or transports, (represented at
16 128-134 in Figure 2), and the operating system.

17 Essentially, NDIS defines a fully abstracted environment for NIC driver
18 development. Thus, for every external function that a NIC driver needs to perform, from
19 registering and intercepting NIC hardware interrupts to communicating with transport
20 protocol drivers to communicating with an underlying NIC via register manipulation and
21 port I/O, it can rely on NDIS APIs to perform the function. To provide this level of
22 abstraction and resulting portability, NDIS uses an export library referred to as the NDIS
23 Interface Library Wrapper (not shown). All interactions between NIC driver and protocol
24 driver, NIC driver and operating system, and NIC driver and NIC are executed via calls to

1 wrapper functions. Thus, instead of writing a transport-specific driver for Windows NT,
2 network vendors provide the NDIS interface as the uppermost layer of a single network
3 driver. Doing so allows any protocol driver to direct its network requests to the network
4 card by calling this interface. Thus, a user can communicate over a TCP/IP network and a
5 DLC (or an NWLINK, or DECnet, VINES, NetBEUI and so forth) network using one
6 network card and a single network driver.

7 At the network and data link layers are transport, protocol and related drivers,
8 shown by way of example in Figure 2 at 128-134. In Windows NT, a transport protocol
9 driver is a software component that implements a transport driver interface (TDI), or
10 possibly another application-specific interface at its upper edge, to provide services to
11 users of the network. In Windows NT, the TDI provides a common interface for
12 networking components that communicate at the Session Layer, such as the Redirector and
13 Server functions illustrated at functional block 138. As is well known, transport protocols
14 act as data organizers for the network, essentially defining how data should be presented to
15 the next receiving layer and packaging the data accordingly. They allocate packets
16 (sometimes referred to in the Windows NT context as NDIS packets), copy data from the
17 sending application into the packet, and send the packets to the lower level device driver
18 by calling NDIS, so that the data can be sent out onto the network via the corresponding
19 NIC.

20 It will be appreciated that additional functions, or tasks, can also be performed on
21 the data packet as it passes through the various network layers, typically at layers 3 and 4
22 of the network model. In accordance with the present invention, such additional functions
23 or task may be performed instead by the NIC hardware.
24

1 For instance, one task conventional performed by transport protocol drivers may be
2 to calculate a checksum value and then append it to the packet. This helps to assure the
3 integrity of the data as it traverses network links. Generally, this operation requires the
4 transport protocol corresponding with the sender of the network packet to append it with a
5 number calculated by adding up the data elements composing the packet. The receiver of
6 the packet then compares the appended checksum number to the data, thereby confirming
7 that data was not changed in transit. This checksum calculation and comparison may
8 instead be offloaded to the NIC hardware.

9 Another related task that could optimally be performed in the NIC hardware is the
10 calculation of a message digest for the data packet. Like the checksum, a message digest
11 is used to guarantee the integrity of the data in the packet. In addition, a message digest
12 can be used to guarantee the authenticity of the data by assuring that the party who sent the
13 message is who they purport to be. Calculation of a message digest is very CPU intensive,
14 and is a function that is expensive to implement in software.

15 Another desirable function is the encryption of the data within the packet.
16 Encryption refers to the cryptographic process of transforming the message in the packet
17 so that it becomes impossible for an unauthorized reader of the packet to actually see the
18 contents of the message without prior knowledge of the encryption key. Of course,
19 cryptographic algorithms also tend to be very CPU and memory intensive, and can be
20 prohibitively expensive if performed in software. Examples of such encryption include
21 Secure Socket Layer protocol (SSL) encryption and Internet security protocol encryption
22 or "IPSec." As is well known, SSL and IPSec encryption protocols are both very CPU and
23 memory intensive.
24

1 Another task that can be performed on the data packet is TCP or UDP
2 segmentation. As is well known, TCP and UDP protocols segment large data packets into
3 segments that align with the maximum data size allowed by the underlying network. For
4 instance, Ethernet allows a maximum of 1514 byte packets on the network. Thus, if TCP
5 or UDP must send 64 Kbytes for example, it must parse the data into 1514 byte segments.

6 In addition, packet classification is often performed on data packets as they are
7 received from a network. Packet classification includes the identification of data packet
8 for Quality of Service (QoS). In other words, each data packet contains fields which
9 define modes of service that should be performed on the data packet in order to ensure
10 optimal performance. For example, data packets that include video and audio may require
11 certain functions to be performed thereon in order to ensure high fidelity in the
12 presentation of video and audio. The classification of the data packets in order to identify
13 the mode of service may be offloaded to the NIC hardware in accordance with the present
14 invention.

15 Packet filtering may also be performed by the NIC hardware instead of by the CPU.
16 In particular, data packets may be evaluated to determine whether they are characteristic of
17 being part of a denial of service attack. Often, server software is created under the
18 assumption that clients that request service are not malicious. However, a client having
19 sufficiently educated, but malicious, user may often issue requests to the server that are
20 designed to prevent the server from servicing other users. Denial of service attack filtering
21 refers to the ability to evaluate data packets to determine if they are characteristic of a
22 denial of service attack.

23 One example of a denial of service attack is a "SYN flood attack," in which, for
24 one reason or another, the client does not send any final acknowledgment to the server's

1 SYNchronize-ACKnowledge (SYN-ACK) response in the handshaking sequence. This
2 causes the server to continue signaling until the server eventually times out.

3 Another type of denial of service attack is called a fragmentation or "teardrop"
4 attack. Internet Protocol requires a packet be fragmented if the packet is too large for
5 handling in the next router. The later fragmented packet identifies an offset to the
6 beginning of the first packet that enables the entire packet to be reassembled at the receive
7 side. In the fragmentation attack, the user puts a counterfeit offset value in the later
8 fragment, often causing the receive side to become dysfunctional. The NIC hardware may
9 be configured to handle filtering for these types of denial of service attacks by evaluating
10 data packets to determine if they are characteristic of a particular kind of denial of service
11 attack.

12 These and other functions are typically performed by the computer CPU 20 in
13 software components residing at the various network layers, and thus can utilize substantial
14 computer resources, resulting in an overall decrease in the computer system performance.
15 Thus, offloading these, or other similar tasks, so that they can instead be performed at the
16 corresponding NIC hardware can greatly increases the overall speed and efficiency of the
17 computer system.

18 As previously noted, the basic unit of data transmission in a Windows NT or
19 similar layered networking model is the data packet. In the Windows NT environment, the
20 data packet is referred to as the NDIS packet. Each packet travels from the top of the stack
21 (*i.e.*, layer 5 in the ISO stack) to the lowest software layer (*i.e.*, layer 2 in the ISO stack).
22 Thus, the packet defines a data structure that is common through each level as it proceeds
23 through the layers during transmission and reception of data. By way of example, Figure 3
24 illustrates the path followed by the packet as it proceeds down through the respective

1 layers to the NIC, shown at 100 as an Ethernet NIC. As noted above, the transport driver
2 128 receives data from a sending application and packages it in packet form consistent
3 with the underlying protocol, and then forwards the packet to the lower level device driver
4 116 via the NDIS interface 126. In addition, the transport protocol may perform other
5 functions on the packet (*e.g.*, checksum calculation, etc.). Alternatively, other functional
6 components may reside in the network layer or data link layers that perform additional
7 functions on the packet, such as the IP Security function 144 (*e.g.*, encryption and/or
8 message digest calculation) illustrated in Figure 3.

9 In one preferred embodiment of the present invention, the data packet 142 is the
10 means by which computing tasks are offloaded to the peripheral device, such as the NIC
11 hardware 100. For instance, in Figure 3 the application data 140 is passed down from the
12 upper layers of the network model to an appropriate transport protocol driver, such as
13 TCP/IP 128. The driver repackages the data into an appropriate data packet 142. Then,
14 depending on whatever additional functions are to be performed on this particular data
15 packet 142 a functional component is included that appends a predefined data structure,
16 referred to as the packet extension, to the data packet. As will be discussed in further
17 detail below, the contents of the packet extension indicate which task, or tasks, are to be
18 performed on the data packet when it reaches the NIC 100. When the data packet 142
19 reaches the network driver 116, the contents of this packet extension are queried by the
20 driver 116 so as to ascertain which task(s) is to be performed by the NIC 100. The driver
21 116 then controls/manipulates the hardware on the NIC so that it will perform whatever
22 functional tasks have been requested via the contents of the packet extension.

23 For example, in Figure 3, the data packet 142 is passed to a software component
24 144, which could be implemented separately or implemented as a part of the transport

1 protocol driver itself, that appends a packet extension to the packet 142. Data will be
2 included within in packet extension depending on the particular task that is to be offloaded.
3 For instance, if an IP security function is to be implemented, data that indicates that the
4 NIC 100 should encrypt the data packet in accordance with a specified encryption key
5 would be included. Of course, the software component 144 could append predefined data
6 such that any one of a number of functions, such as those discussed above, would be
7 performed at the hardware level instead of by software components that reside in the
8 network layers. The device driver 116 will extract the information from the packet
9 extension, and then invoke the specified task(s) at the NIC 100.

10 Figure 4 illustrates one presently preferred embodiment of the general structure of
11 the data packet 142. While the packet 142 can be of any format depending on the exact
12 network environment being used, in a Windows NT environment the packet is formatted
13 according to NDIS, and includes information such as a packet descriptor; flags whose
14 meaning is defined by a cooperating device driver(s) and protocol driver(s); areas for
15 storage of Out-of-band data (OOB) associated with the packet; information relating to the
16 length of the packet; and pointers to memory locations relating to the data content of the
17 packet.

18 Figure 4 further illustrates the additional data structure field that is appended to the
19 NDIS data packet to identify task offloads -- the packet extension 150. As discussed, it is
20 this packet extension 150 which defines a data structure containing information necessary
21 for the identification of the particular task, or tasks, that are being offloaded to the
22 destination NIC. In the preferred embodiment, for each task offload type (*e.g.*, checksum,
23 encryption/decryption, etc) a predefined data field will be included within the packet
24 extension 150. This data field can simply be in the form of a control flag or flags, which

1 merely indicates that a particular function be performed (such as a checksum), or the
2 information can be in the form of a pointer to a data structure that further defines how a
3 task should be carried out.

4 The NIC 100 may be configured to recognize the task offload control field in the
5 packet extension 150 as applying only to the attached packet. Thus, for example, if a
6 particular packet contains a flag signifying the NIC 100 is to perform a checksum
7 operation, the NIC 100 will perform a checksum operation on the attached packet only.
8 However, if there is no such flag for a given packet, then the NIC 100 does not perform the
9 checksum operation for that packet. Alternatively, the task offload control field in the
10 packet extension 150 may indicate that until the NIC 100 is instructed otherwise, the NIC
11 is to perform the offloaded task on this and all subsequent packets sent out over the
12 network.

13 The packet extension 150 may also instruct the NIC station at the receive station of
14 what tasks to perform. For example, the packet extension may instruct the receive station
15 NIC that it is to perform certain functions appropriate when receiving a packet from a
16 network such as, for example, decryption, checksum operations, packet classification,
17 packet filtering to guard against denial of service attacks, packet reassembly, and any other
18 receive function that an NIC may be configured to perform. Of course, the sending station
19 may have no idea on the task offload capabilities of the receive station NIC. If the receive
20 station NIC is unable to perform the requested function, then receive station CPU will
21 accomplish the function instead. Thus, the sending station may have some control over
22 what the NIC at the receiving station does to a particular packet.

23 In addition, the sending station may also use packet extensions to control the NIC
24 100 when it functions to receive packets. For example, a particular packet to be sent out

1 over the network may include a packet extension that includes an instruction to the NIC
2 100 that when it receives packets, it is to perform decryption of the packet. In this case,
3 the instruction does not apply to the packet to be sent out over the network since the
4 instruction is for performing functions to packets received from the network. Therefore, in
5 this case, the packet extension for a packet to be sent over the network serves as a
6 mechanism to control what receive functions are offloaded to the NIC 100 when receiving
7 packets. The sending station has an opportunity to control the NIC 100 receive functions
8 whenever it sends a packet, on a packet-by-packet basis.

9 Referring again to the example shown in Figure 4, the packet extension 150
10 contains a number of data fields 151 to 153 that control the sending functions offloaded to
11 an NIC (e.g., NIC 100) and number of data fields 154 to 158 that control the receiving
12 functions offloaded to an NIC (e.g., NIC 100 or an NIC of a receiving station).

13 For example, the packet extension 150 includes a data field 151 signifying that the
14 NIC 100 perform a checksum operation. This indicates to the sending NIC 100 that the
15 sending NIC 100 itself is to perform a checksum operation on the appended packet. By
16 way of example and not limitation, a preferred embodiment of the packet extension data
17 structure for signifying that the NIC perform a checksum operation has the following
18 structure:

```
19  
20 typedef struct _NDIS_TCP_IP_CHECKSUM_PACKET_INFO  
21 {  
22     union  
23     {  
24         struct  
25         {  
26             ULONG NdisPacketChecksumV4:1;  
27             ULONG NdisPacketChecksumV6:1;
```

```

1      }
2      Transmit;
3      struct
4      {
5          ULONG    NdisPacketTcpChecksumFailed:1;
6          ULONG    NdisPacketUdpChecksumFailed:1;
7          ULONG    NdisPacketIpChecksumFailed:1;
8          ULONG    NdisPacketTcpChecksumSucceeded:1;
9          ULONG    NdisPacketUdpChecksumSucceeded:1;
10         ULONG    NdisPacketIpChecksumSucceeded:1;
11         ULONG    NdisPacketLoopback:1;
12     }
13     Receive;
14
15     ULONG    Value;
16 };
17 }
18
19     NDIS_TCP_IP_CHECKSUM_PACKET_INFO,
20     *PNDIS_TCP_IP_CHECKSUM_PACKET_INFO;

```

In this particular packet extension data structure, if the variables *NdisPacketChecksumV4* and *NdisPacketChecksumV6* both are not set, then the device driver should send the data packet without doing any checksum on it.

The packet extension 150 also may include a security function data field 152, to indicate that the NIC 100 should perform security features such as would be performed in connection with an SSL or IPSec encryption of packet data and/or the calculation of a message digest. For this type of security task, field 152 preferably contains a pointer to a memory location containing a data structure (e.g., data structure 160), which in turn contains information relevant to the performance of the encryption and/or message digest functions. Under some circumstances, the inclusion of a pointer to a memory location having pertinent data has advantages over storing actual data within the packet extension itself.

One such advantage is illustrated in Figure 5, which illustrates an example of a preferred program sequence that corresponds to the situation where multiple successive

1 data packets may need to have the same type of encryption or digest calculation operation
2 performed. After beginning at program step 302, program step 304 determines whether the
3 current packet is the first packet in the sequence of packets. If so, then this first packet is
4 provided with information, or context, that is to be used for that operation on the
5 succeeding packets as well. For instance, the first packet will have a packet extension that
6 sets forth the particular encryption key to use. This value, or context, will be stored in a
7 separate memory location, or handle. Successive packets will not have to include this
8 information, and would, for instance, only have provide a pointer to the memory location
9 where the encryption key (or other context information) is stored, as is denoted at program
10 step 308. This approach reduces the overall size of subsequent data packets in the
11 sequence of packets, and further enhances the efficiency and portability of the task offload
12 method.

13 The packet extension 150 may also include a segmentation data field 153 (either for
14 TCP or UDP packet segmentation), which instructs the NIC 100 to perform segmentation
15 of bulk data into constituent packets. In this case, the packet extension 150 may be
16 appended to bulk data instead of a packet. The flags in the packet extension may then be
17 seen to apply to all of the resulting packets.

18 In addition to these send functions performed on packets sent out by the NIC 100,
19 the NIC 100 may be instructed to perform certain offloaded receive functions when
20 receiving packets from the network. These receive function instructions may be given in
21 the packet extension 150 even though the appended packet is sent out over the network,
22 and thus the NIC 100 does not perform the receive function on the appended packet.
23 Examples of such offloadable receive functions are illustrated by the data fields 154
24

1 through 158. Some of these offloadable receive functions closely correspond to
2 offloadable send functions. Others have no close corresponding send function.

3 Referring to Figure 4, a checksum data field 154 indicates that for all received
4 packets for which checksum calculation has been performed at the send station, the NIC
5 100 is to perform a checksum calculation to make sure that the data has not been altered en
6 route to the NIC 100. This checksum field 154 may control the checksum receive function
7 for all packets indefinitely or for a given time period unless expressly contravened by a
8 subsequent checksum field.

9 The packet extension may also include a security flag 155 which indicates that the
10 NIC 100 should perform security features as packets are received such as would be
11 performed in connection with an SSL or IPSec decryption of packet data and/or the
12 calculation of a message digest. For this type of security task, field 155 preferably
13 contains a pointer to a memory location containing a data structure (e.g., data structure
14 160), which in turn contains information relevant to the performance of the encryption
15 and/or message digest functions.

16 The packet extension may also include an assembly data field 156 which instructs
17 the NIC 100 to assemble the received packets into bulk data, a classification field 157
18 which instructs the NIC 100 to classify each packet for quality of service as described
19 above, and a Denial of Service (DOS) attack filter field 158 for filtering incoming packets
20 for characteristics of a DOS attack as described above. The packet extension 150 may also
21 include receive function data fields similar to flags 154-158 intended for the receiving
22 station NIC to perform on the attacked packet.

23 In a preferred embodiment, the information contained within the packet extension
24 150 is queried by the particular device driver to which the packet 142 is sent. In the

1 Windows NT environment described in the illustrated embodiments, this type of function
2 would preferably be performed by making appropriate NDIS function calls. For instance,
3 a call to a predefined NDIS function that returns a pointer to the packet extension 150
4 memory location for the packet could be performed. The device driver software could then
5 identify which tasks are to be performed and, depending on the task(s) offloaded,
6 operate/manipulate the driver's corresponding NIC hardware in the appropriate manner.

7 Utilizing the actual data packet to offload computing tasks from the computer
8 processor to the hardware peripheral is advantageous for a number of reasons. For
9 example, the transport driver can utilize the capabilities of the peripheral on a packet-by-
10 packet basis. This allows tasks to be downloaded dynamically, and the capabilities of a
11 peripheral can be used on an as-needed basis. Thus, if the processing overhead for the
12 computer system is low at a particular point in time, it may be desirable to perform certain
13 tasks on the computer processor in a conventional fashion. Alternatively, if CPU is
14 heavily loaded with other computing tasks, then it can offload tasks to peripheral devices
15 by merely appending the requisite packet extension to the data packets.

16 Another advantage is the ability offload multiple tasks by way of a single packet,
17 and essentially "batch" a number of operations at once. For instance, when the computer
18 processor performs a checksum operation, or an encryption operation, the entire data field
19 must be loaded into a memory location before the operation can be completed, *i.e.*, the
20 calculation of the checksum or the encryption of the packet data. Moreover, because of the
21 layered networking model only one operation can be performed at a time, thereby requiring
22 the data to be copied into memory multiple times. However, the per-packet approach
23 allows multiple tasks to be offloaded in one packet. Thus, the hardware peripheral can
24 perform two or more operations in a single pass on the data, depending on the capabilities

1 of the hardware, and thereby greatly increase the throughput and efficiency of the
2 computer system.

3 It will be appreciated that while the above scheme is particularly useful in
4 connection with the ability to specify tasks to be offloaded to a particular NIC, the per-
5 packet transfer of information could be used in other ways as well. For instance, if a
6 particular NIC is capable of scheduling the delivery of packets at predetermined times, the
7 packet extension data structure could be used to pass information identifying how and/or
8 when the packet should be sent by the NIC hardware.

9 In a preferred embodiment of the present invention, before a transport protocol
10 driver will append a packet extension to a data packet so as to offload a particular task to a
11 NIC, two additional functions are initially performed. Insofar as there are many different
12 types of hardware peripherals, each having varying processing capabilities, embodiments
13 of the present invention preferably provide a means by which the transport driver can first
14 query the task offload capabilities of the peripherals that are connected to the computer
15 system. Once these capabilities have been ascertained, the transport protocol driver can
16 then set, or enable, those tasks that are of interest. Once enabled, the specified tasks can be
17 subsequently utilized on a per-packet basis in the manner described above.

18 Figure 6 illustrates one presently preferred set of program steps for implementing
19 this ability to query the task offload capabilities of a peripheral, such as a NIC, and for then
20 setting/enabling those tasks that may be needed. Preferably, the illustrated program steps
21 are implemented as a software component or module that is integrated within a transport
22 protocol driver. Moreover, if implemented in a Windows NT environment, many of the
23 functional operations and inter-driver communications are preferably carried out via the
24 NDIS interface. Of course, if implemented in a different operating system environment, or

1 implemented in connection with non-network types of peripherals, then the illustrated
2 program steps would have to be modified accordingly.

3 Beginning at program step 202, executable program instructions are performed by
4 the computer system CPU (e.g. processing unit 21 in Figure 1) that query the task offload
5 capabilities of a peripheral(s) connected to the computer system. In the exemplary
6 environment, the peripheral(s) would be the NIC(s) attached to the system. Preferably,
7 each NIC device driver (116-120 in Figure 2) will have associated with it a predefined task
8 offload buffer location(s) that each contains the task offload capabilities for that device
9 driver and its corresponding NIC. At a minimum, the task offload buffer will identify the
10 particular task(s) supported by the NIC and its device driver, and will also include any
11 information specific to each individual task supported. In the preferred embodiment, the
12 contents of a device driver's task offload buffer are retrieved by way of an NDIS function
13 call.

14 By way of example, and not limitation, a preferred embodiment of the driver's task
15 offload buffer has the following structure:

```
17 typedef struct _NDIS_TASK_OFFLOAD
18 {
19     ULONG          Size;
20     NDIS_TASK      Task;
21     ULONG          OffsetNextTask;
22     ULONG          TaskBufferLength;
23     UCHAR          TaskBuffer[1];
24 }
25 NDIS_TASK_OFFLOAD,
26 *PNDIS_TASK_OFFLOAD;
```

27 wherein: *Size* This variable is set to the size of this structure, and is
28 used to determine the version of the structure
29 in use.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

<i>Task</i>	This indicates the type of task offload defined in the structure.
<i>OffsetNextTask</i>	Offset to the next task offload buffer. The value is 0 if this is the last task supported by this driver/NIC.
<i>TaskBufferLength</i>	Length of the task offload buffer that follows this structure.
<i>TaskBuffer</i>	This is a task specific buffer that contains information/data necessary for describing the particular task offload defined in this structure.

Once the particular task offload capabilities of each peripheral have been ascertained, processor 21 proceeds to the computer executable instructions corresponding to the function illustrated at program step 204 in Figure 6. Here, it is determined whether any of the particular task offload capabilities supported by each peripheral NIC are of any interest to the transport. If no tasks are supported by the peripheral, or if the tasks that are supported are not useful to the transport, then the processor will proceed directly to program step 208, where this particular line of processing stops, and the transport proceeds in its normal fashion. Alternatively, if one or more of the tasks identified are of interest, then program step 206 is performed. At this step, executable instructions are performed that set those particular task offload capabilities the protocol driver wants enabled. This informs the device driver and its corresponding NIC what type of task offloading it may expect on a per packet basis. In a preferred embodiment, the task offload capability is set by a setting appropriate data values in a data structure that is then passed to the corresponding device driver via an NDIS interface function call. For example, for a particular task, the appropriate bits in the task's task offload buffer structure can be set by the transport, thereby enabling the task for that driver/NIC. The transport can thus enable

1 any number of task offload capabilities of each NIC/NIC driver that it may want to later
2 use.

3 Once the desired offload capabilities have been enabled for each device driver and
4 its corresponding NIC, the computer system processor 21 proceeds to program step 208,
5 where processing for this particular function ends. At this point, the transport driver can
6 utilize each of the task offload capabilities that have been enabled on a per packet basis in
7 the manner previously described.

8 In summary, embodiments of the present invention provide distinct advantages
9 over what is currently available in the prior art. Specific processing tasks that would
10 otherwise be performed in a computer system's processor and memory, are instead
11 downloaded to a particular peripheral device, or devices, that are connected to the
12 computer. The computing task is then performed by the peripheral, thereby saving
13 computer system resources for other computing tasks. Especially where the offloaded
14 processing task is CPU and/or memory intensive, this task offload scheme can dramatically
15 increase the overall computing efficiency of the computer system. Advantageously, the
16 disclosed method for offloading computing tasks provides a means by which tasks can be
17 offloaded on a dynamic, as-needed basis. As such, the processor is able to offload tasks in
18 instances where it is busy processing other computing tasks and processor overhead is
19 high. Alternatively, when demand on the computer system's processing resources is
20 lower, the processor may instead perform the task on its own. In addition, multiple tasks
21 can be offloaded in batches to a particular peripheral. Oftentimes, a peripheral will be
22 optimized so as to perform such multiple tasks in a much more efficient manner than the
23 computer processor -- again resulting in improvements to the overall system's efficiency.
24 Finally, a method of the present invention provides a processing scheme by which the

1 particular task offload capabilities of a peripheral can be queried, and thereafter selectively
2 enabled. In this way, the computer system can easily discern the capabilities of its various
3 peripheral devices, and take advantage only of those processing capabilities that are, or
4 may be, thereafter needed.

5 The present invention may be embodied in other specific forms without departing
6 from its spirit or essential characteristics. The described embodiments are to be considered
7 in all respects only as illustrated and not restrictive. The scope of the invention is,
8 therefore, indicated by the appended claims rather than by the foregoing description. All
9 changes which come within the meaning and range of equivalency of the claims are to be
10 embraced within their scope.

11 What is claimed and desired to be secured by United States Letters Patent is:
12
13
14
15
16
17
18
19
20
21
22
23
24